O'REILLY®

bhv®

Обработка данных на Python. Data Wrangling и Data Quality

Начните работу с чтения, очистки и анализа данных

+ Материалы на www.bhv.ru

Сьюзен Макгрегор

Сьюзен Макгрегор

Обработка данных на Python. Data Wrangling и Data Quality

Начните работу с чтения, очистки и анализа данных

Санкт-Петербург
«БХВ-Петербург»
2024

УДК 004.43
ББК 32.973.26-018.1
М15

Макгрегор С.

М15 Обработка данных на Python. Data Wrangling и Data Quality: Пер. с англ. — СПб.: БХВ-Петербург, 2024. — 432 с.: ил.

ISBN 978-5-9775-1846-8

Книга посвящена первичной обработке данных (Data Wrangling) на Python и оценке их качества (Data Quality). Материал содержит основополагающие концепции, экспертные советы и ресурсы, необходимые для первичной обработки, извлечения, оценки и анализа данных. Все темы раскрыты на простых и наглядных примерах из практики. Даны необходимые и достаточные сведения о языке программирования Python 3.8+ для чтения, записи и преобразования данных из различных источников, а также для обработки данных в больших масштабах. Приведены лучшие практики документирования и структурирования кода. Описан сбор данных из файлов, веб-страниц и API. Рассмотрены приемы проведения базового статистического анализа наборов данных, а также наглядные и убедительные способы визуализации и представления данных. Изложение рассчитано как на новичков по обработке данных, так и на профессионалов.

Электронный архив на сайте издательства содержит цветные иллюстрации к книге.

Для специалистов по обработке данных

УДК 004.43
ББК 32.973.26-018.1

Научный редактор:

Архитектор решений, IT-компания «Яндекс» *Дмитрий Бардин*

Группа подготовки издания:

Руководитель проекта	*Евгений Рыбаков*
Зав. редакцией	*Людмила Гауль*
Перевод с английского	*Михаила Райтмана*
Редактор	*Наталья Смирнова*
Компьютерная верстка	*Натальи Смирновой*
Оформление обложки	*Зои Канторович*

Подписано в печать 06.12.23.
Формат 70×100$^{1}/_{16}$. Печать офсетная. Усл. печ. л. 34,83.
Тираж 1500 экз. Заказ № 8227.
"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

Отпечатано с готового оригинал-макета
ООО "Принт-М", 142300, М.О., г. Чехов, ул. Полиграфистов, д. 1

ISBN 978-1-492-09150-9 (англ.)
ISBN 978-5-9775-1846-8 (рус.)

Содержание

Введение **11**

Для кого предназначена эта книга? 12
Поедете сами или возьмете такси? 12
Кому не следует читать эту книгу? 13
Что ожидать от этой книги 13
Типографские соглашения 14
Использование примеров кода 15
Цветные иллюстрации 16
Возможности онлайнового обучения от компании O'Reilly 16
Как связаться с нами 16
Благодарности 17

ГЛАВА 1. Введение в первичную обработку (выпас) и качество данных **19**

Что такое выпас данных? 20
Что такое качество данных? 22
    Целостность данных 23
    Соответствие данных 23
Почему мы выбрали Python? 25
    Универсальность 25
    Доступность 25
    Удобочитаемость 26
    Сообщество 26
Альтернативы языку Python 27
Написание и исполнение программ на языке Python 27
Работа с кодом Python на локальном устройстве 30
    Введение в работу с командной строкой 30
    Установка языка Python, среды Jupyter Notebook и редактора кода 33
Работа с кодом Python в режиме онлайн 39
Hello World! 39
    Создание файла автономного сценария Python при помощи Atom 39
    Создание блокнота кода Python в Jupyter Notebook 41
    Создание блокнота кода Python в Google Colab 42
Создаем программу 43
    В файле автономного сценария 43
    В блокноте 43

Исполняем программу ... 43
В файле автономного сценария ... 43
В блокноте ... 44
Документирование, сохранение и управление версиями своего кода ... 44
Документирование ... 44
Сохранение ... 46
Управление версиями ... 46
Заключение ... 56

ГЛАВА 2. Введение в Python ... 57

"Части речи" языков программирования ... 58
Существительные ≈ переменные ... 59
Важно ли конкретное имя? ... 61
Наилучшие практики для именования переменных ... 62
Глаголы ≈ функциям ... 62
Применение пользовательских функций ... 67
Библиотеки: занимаем пользовательские функции у других программистов ... 68
Структуры управления: циклы и условные операторы ... 69
Циклы ... 69
Условные операторы ... 72
Понимание ошибок ... 77
Синтаксические ошибки ... 78
Ошибки времени исполнения ... 79
Логические ошибки ... 83
Отправляемся в путь с данными Citi Bike ... 85
Начинаем с создания псевдокода ... 86
Масштабирование ... 92
Заключение ... 93

ГЛАВА 3. Понимание качества данных ... 95

Оценка соответствия данных ... 97
Достоверность данных ... 98
Надежность данных ... 100
Репрезентативность данных ... 101
Оценка целостности данных ... 104
Необходимые, но недостаточные ... 106
Важные ... 108
Достижимость ... 111
Улучшение качества данных ... 114
Очистка данных ... 115
Аугментация данных ... 115
Заключение ... 116

ГЛАВА 4. Работа с файловыми и канальными данными на языке Python ... 117

Структурированные и неструктурированные данные ... 119
Работа со структурированными данными ... 123
Файловые табличные форматы данных ... 124
Выпас табличных данных посредством языка Python ... 126
Выпас реальных данных: понимание безработицы ... 133
XLSX, ODS и все остальные ... 136
Данные фиксированной ширины ... 143
Канальные данные — интерактивные обновления через Интернет ... 147
Выпас канальных данных средствами языка Python ... 150
Формат JSON: данные следующего поколения ... 160
Работа с неструктурированными данными ... 165
Текст на основе изображений: доступ к данным в формате PDF ... 165
Выпас PDF-данных, используя Python ... 166
Обращение к таблицам PDF посредством Tabula ... 171
Заключение ... 171

ГЛАВА 5. Доступ к интернет-данным ... 173

Доступ к веб-данным XML и JSON ... 175
Знакомство с API-интерфейсами ... 178
Базовые API-интерфейсы на примере поисковой системы ... 179
Специализированные API-интерфейсы: добавление простой аутентификации ... 181
Получение ключа для API-интерфейса FRED ... 181
Использование ключа API для запроса данных ... 182
Чтение документации по API-интерфейсу ... 183
Защита своего ключа API при использовании сценариев Python ... 186
Создание файла учетных данных ... 188
Использование учетных данных в отдельном сценарии ... 189
Основы работы с файлом .gitignore ... 190
Специализированные API-интерфейсы: работа с протоколом OAuth ... 193
Получение учетной записи разработчика Twitter ... 194
Создание приложения и учетных данных Twitter ... 196
Кодирование ключа API и ключа секрета API ... 201
Запрос токена доступа и данных из API-интерфейса Twitter ... 202
Этические нормы при работе с API-интерфейсами ... 206
Извлечение веб-данных: источник данных последней надежды ... 207
Осторожно извлекаем данные с веб-сайта УГПТ ... 210
Использование средств инспектирования браузера ... 211
Решение Python для извлечения данных из веб-страницы: библиотека Beautiful Soup ... 214
Заключение ... 218

ГЛАВА 6. Оценка качества данных 219

Пандемия и программа PPP 221
Оценка целостности данных 222
Имеют ли данные известное происхождение? 223
Актуальны ли данные? 223
Полные ли данные? 224
Хорошо ли данные аннотированы? 236
Являются ли данные крупномасштабными? 242
Непротиворечивы ли данные? 244
Многомерны ли наши данные? 248
Атомарны ли данные? 250
Понятны ли данные? 250
Размерностно структурированы ли данные? 252
Оценка соответствия данных 253
Достоверность данных 253
Надежность данных 257
Репрезентативность данных 258
Заключение 259

ГЛАВА 7. Очистка, преобразование и дополнение данных 261

Выбор подмножества данных системы Citi Bike 262
Простое разбиение 263
Регулярные выражения: супермощное средство сопоставления строк 265
Создание дат 270
Удаление хлама из файлов данных 272
Декодирование дат Excel 276
Создание настоящих данных CSV из данных фиксированной ширины 279
Исправление разнообразности написаний 282
Тернистый путь к "простым" решениям 288
Опасные подводные камни 290
Дополнение данных 292
Заключение 294

ГЛАВА 8. Структурирование и рефакторинг кода 296

Обзор пользовательских функций 296
Многократное использование кода 297
Аккуратное и понятное документирование 297
Недостаточная функциональность по умолчанию 298
Область видимости 298
Определение параметров функции 301
Доступные опции 302
Предоставление аргументов 303
Возвращаемые значения 303
Работа со стеком 305

Рефакторинг для получения удовольствия и прибыли ... 306
Функция для определения рабочих дней ... 306
Опрятные метаданные ... 309
Использование pydoc для документирования сценариев и пользовательских функций ... 317
О полезности аргументов командной строки ... 321
Отличия во взаимодействии со сценариями в автономных файлах и блокнотах ... 325
Заключение ... 325

ГЛАВА 9. Введение в анализ данных ... 327

Вся суть — в контексте ... 328
Одинаковые, но не совсем ... 329
Что типично? Оценка центральной тенденции ... 329
Что это значит? ... 330
Поразмыслим нестандартно: выявляем выбросы ... 332
Визуализация для анализа данных ... 332
Какова форма наших данных? Учимся понимать гистограммы ... 336
Вопрос за $2 миллиона ... 346
Пропорциональный ответ ... 359
Заключение ... 362

ГЛАВА 10. Представление данных ... 364

Основы визуального красноречия ... 365
Сформулируйте свои данные ... 367
Диаграммы, графики и картограммы — вот это да! ... 368
Круговые диаграммы ... 369
Линейчатые и столбчатые диаграммы ... 372
Линейные диаграммы ... 377
Диаграмма рассеяния ... 380
Картограммы ... 383
Элементы красноречивых визуальных эффектов ... 386
"Мелкие" детали действительно имеют значение ... 386
Доверяйте своим глазам (и экспертам) ... 387
Выбор масштаба ... 388
Выбор цветовой гаммы ... 389
Прежде всего делайте аннотации! ... 389
От базового к красивому: настройка визуализации с помощью seaborn и matplotlib ... 390
Выйдите за рамки основ ... 395
Заключение ... 396

ГЛАВА 11. За пределами Python ... 397

Дополнительные инструменты для анализа данных ... 398
Программы для работы с электронными таблицами ... 398
OpenRefine ... 399

Дополнительные инструменты для обмена и представления данных ... 402
Редактирование изображений в форматах JPG, PNG и GIF ... 402
Программное обеспечение для редактирования SVG и других векторных форматов ... 402
Размышления об этике ... 404
Заключение ... 405

Приложение А. Другие ресурсы по программированию на Python ... 406

Официальная документация Python ... 406
Установка ресурсов Python ... 407
Где искать библиотеки ... 407
Следите за остротой своих инструментов ... 408
Где получить больше информации ... 409

Приложение Б. Еще несколько слов о Git ... 410

Вы запускаете команду git push/pull и оказываетесь в странном текстовом редакторе ... 410
Ваша команда git push/pull отклоняется ... 412
Выполните команду git pull ... 412
Краткое руководство по Git ... 414

Приложение В. Поиск данных ... 416

Репозитории данных и API ... 416
Эксперты по предметным вопросам ... 417
Запросы FOIA/L ... 418
Кастомные методы сбора данных ... 419

Приложение Г. Ресурсы для визуализации и информационного дизайна ... 421

Основополагающие книги по визуализации информации ... 421
Краткое руководство, за которым вы потянетесь ... 422
Источники вдохновения ... 422

Об авторе ... 423

Колофон ... 424

Предметный указатель ... 425